Methods of Selecting Variables in Data Disclosure

Charles C Lin[1] and Tom Garin[1]

[1]Data Governance and Analytics, Department of Veterans Affairs
810 Vermont Avenue NW, Washington, DC 20420

Key words: Re-identification, Variable selection, Data disclosure, Data sparseness

Abstract

Data released for public use needs to be protected for personal identified information such as name and Social Security Number and any unusual pattern in data which could have a high risk of re-identification. All privacy protected information should be stripped from released variables in data according to a set of disclosure guidelines. Excessive restrictions of released data elements, however, could render the data useless. The task of proper data disclosure with minimum risk of re-identification can be overwhelming when the number of variables becomes large. The brute force approach of examining all becomes infeasible as the number of possible combinations of variables increases exponentially. To prepare a data for public use, this study proposes systematic methods examining each variable with respect to both risks and benefits. These methods compare a subset of variables to another and traverse back and forth from subsets to subsets in a step-by-step manner until a stopping criterion is met. The final subsets of variables will provide reasonable choices of variables for public release. The methodology with its implementation in SAS program will be discussed in details and demonstrated with examples.

Introduction

Data disclosure to public has been a general practice in government and public for many years upon requests of Freedom of Information Act (FOIA) signed into law in 1966 and recently issued Open Data guidance to all Federal agencies starting May 2013, followed Executive Order 13642 of "Making Open and Machine Readable the New Default for Government Information" and OMB Memorandum 13-13 of "Open Data Policy-Managing Information as an Asset". However, releasing data to public has not been an easy task as personally identifiable information (PII) is protected by privacy laws, like the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule in 1996 and Executive Order 13556 for Controlled Unclassified Information (CUI) in 2010. Privacy protection is also facing challenge now as vast amount of information on individuals already are abundant in public domains like vehicle and voter registration records. Sweeney showed more than 87% of US citizens can be identified by the three variables gender, date of birth, and zip code available in public databases with the 1990 Census data [5]. Data can be aggregated to remove PII but will make data no longer granular enough to be fully useful. Balancing privacy against data release requires federal agencies to develop guidelines for reliably de-identifying datasets so they can be released without violating individual's privacy.

Risk of re-identification, the process used to identify individuals connected with information, typically arises when a small count of unusual pattern in values of released variables can be used to identify individuals. Restricting release of variables will certainly reduce occurrences of unusual patterns in data in public and hence lower re-identifying risk. Descriptive variables such as age and gender are common in disclosed data. But many other variables are deemed minimum necessary for release purpose. Pre-screening variables could drastically reduce the number of variables needed for release. Unique patterns in data post high risks of re-identification if the uniqueness is not caused by lack of sufficient data. Restricting variables for release will subsequently lessen number of unique patterns in data.

Identifying uniqueness with relevant variables can be a good practice exploiting sources of high risk of re-identification in data [2][3], but the search should go beyond uniqueness and extend to unusual patterns of small counts which possess high risks as well [1]. Various measures of risk of re-identification associated with a set of variables have been proposed but involved complex computations [3], simpler risk metrics given by Emam [4] are found more applicable. Thus, sets of variables with low values in the risk metrics will then be desired for release purpose.

The task of searching sets of variables with low risk measure can be overwhelming if the number of variables is large. Thus, an automatic process of selecting subsets of variables within a limit in risk is warranted. The proposed three methods select subsets of variables according to criteria derived from the risk metrics of Emam. These methods offer good choices of a set of variables for release without going through all possible subsets of variables and the subsets generated do not require to be of minimum risk. With an implementation of SAS programs, these variables selection methods are illustrated with examples and further discussed in details.

Data for Release

To prepare data for release all variables need to be screened according to disclosure guidelines for privacy protection. Masking or generalization maybe required to hide individual identity such as using age groups or first three digits of zip codes. The HIPPA Safe Harbor known rules includes limiting use of cross-year longitudinal data and across datasets numbers of records, suppressing geographic information below state level, and aggregating ages over 89 into a single age category. [3] [6]

Released Data as Sample of Population

The released data is considered to be a sample of a known population where an attacker can link an individual in the population to the released data by their common variables. Therefore, the focus is to restrict the release of variables in the sample which could reveal unusual values, as combining more unusual values of many variables the re-identifying risk is greatly amplified.

The key variables, such as age and gender in the sample gives a limit number of distinct values, can be used to re-identify a target individual. A set of key variables provides a limit number of patterns of values which defines the universe for re-identification. Each pattern of values can be visualized as a cell in the cross-tabulation of the key variables. The cell size is the number of repetitions of a pattern in the released data/sample and all cell sizes will sum up to the sample size. Thus, cells of smaller cell sizes, with unusual pattern of values, pose greater risks of re-identification. The goal of de-identification is to restrict the release of key variables so that to produce a relatively small number of cells with small cell sizes.

Sample Unique and Population Unique

Sample unique and population unique are cells of cell size one in the sample and population, respectively. Sample unique implies a probable population unique, which poses as the highest risk of re-identification. [2][3] In general, small cell sizes like less than 3 in the sample are considered at a high risk of re-identification [1]. Assuming sample unique is population unique makes computation feasible but leads to conservative estimate.

Methodology of Variable Selection

The development of methodology of variable selection would first rely on assessing the risk of re-identification of released variables in data. The risk is usually measured with limitations as the population for the released data may not be known. By adapting simple risk metrics and with necessary assumptions, we proposed three variable selection methods; the Forward, Backward/Elimination, and Stepwise methods. The Stepwise method is a variation of the Forward method with extra protection on risk tolerance. All methods will provide reasonable choices of released variables within a tolerated limit on risk.

Risk Measures of Re-identification

The Risk Proportion (RP) is the proportion of records with sample cell size less than a predetermined number. The Cell Ratio (CR) is the ratio of the number of non-empty sample cells over the total number of records. Assuming a target individual is known included in the released data, RP and CR will be the same as the risk metrics by Emam, assessing the proportion of records that have a re-identification probability larger than a threshold and the average probability of re-identification, respectively [4]. In general, RP over-estimates (to be conservative) the proportion of population at high risk of re-identification and CR over-estimates the average proportion of population re-identified, the former over-estimation tends to be larger the latter.

Both RP and CR measure the risk of re-identification for a set of key variables. The values of RP and CR will be increased when adding a new key variable, as cells are broken up and produce more cells with small cell sizes and more nonempty cells. Dropping a key variable, or regrouping the values of a key variable such as using 10-year age group instead of age, will collapse cells into less but larger cells and therefore decrease the values of RP and CR. Low values of RP and CR indicate low risk of re-identification. A low value of CR also shows data become sparse due to a small number of distinct patterns for a set of key variables in data. A high value of CR should be preferred among those subsets with the similar number of key variables and their RP values as less sparseness increases the usefulness of released data. Among possible choices of key variables or regrouping for key variables, a set of key variables with a relatively high value of CR among those with low values of RP should then be desired.

Criterion of Selecting Key Variables

The possible choices of a subset of key variables grow exponentially with the number of key variables available which makes it infeasible to evaluate all subsets. But, from a subset of key variables, it's desirable to add a new key variable to increase less in the value of RP and more in the value of CR, or to result to a slower rate of increase in the value of RP relatively to the rate of increase in the value of CR. Thus, a small ratio α of the rate of increase in RP over the rate of increase in CR is desired, where

$$\alpha = \frac{{}^{RP}/_{RP'}}{{}^{CR}/_{CR'}} = \frac{{}^{RP}/_{CR}}{{}^{RP'}/_{CR'}} \quad (1)$$

RP' and CR' are RP and CR of the subset of key variables before a new key variable is added, respectively. The equation shows a small α is equivalent to a small ratio of $\frac{RP}{CR}$. Conversely, removing a key variable which would have produced a large α or $\frac{RP}{CR}$ is preferable.

Stopping Criterion

Starting from a subset of key variables, more key variables can be added subsequently until the risk of re-identification becomes too high. Conversely, the process of succeeding removing key variables should be stopped when the risk can be tolerated. The value of RP between 0 and 1 of a subset of key variables will be used to judge the risk of identification for the subset.

Forward Method

1. Determine all key variables available for disclosure
2. Start by including the key variables known to be released such as gender and age
3. Add the new key variable with the smallest α
4. Stop when the value of RP of the subset of key variables is larger than a pre-set percentage, otherwise continue to Step 3

Backward/Elimination Method

1. Determine all key variables available for disclosure
2. Start by including all key variables
3. Remove the key variable, other than those known to be released such as age and gender, with the largest α
4. Stop when the value of RP of the subset of key variables is smaller than a pre-set percentage, otherwise continue to Step 3

Stepwise Method

1. Determine all key variables available for disclosure
2. Start by including the key variables known to be released such as gender and age

3. Remove the key variable, other than those known to be released such as age and gender, with the largest α until the value of RP is smaller than a pre-set percentage. Then, add the new key variable with the smallest α
4. Stop when the value of RP of the subset of key variables is larger than a pre-set percentage, otherwise continue to Step 3

The Stepwise method is like the Forward method, but will check if to remove any variable from the set of current key variables before adding a key variable. The Stepwise method assures the selected key variables also meet the requirement in the Backward method. To prevent an infinite loop, the stopping percentage for removing a variable should be sufficiently larger than the one for adding a variable and also prohibit removing a variable just added in the last step.

Demonstration of Software Implementation by Examples

The examples showed the results of the software implementation of the methods written in SAS macros. The data is a sample of VA healthcare records with the 16 key variables *sex, ag8r, source, pow, ms, psx, aor, homstate, means, nsurg, nbs, cp, lsr, ethnic, race1,* and *psrcd*. The age group variable, *ag8r*, and variable *sex* were decided to be included in the set of released variables. The Forward method selects the variable *pow* with the smallest α along with the variables *sex* and *ag8r* to be included in Step 1.

Forward step 1:

Varlist: sex ag8r
pow (smallest alpha=0.000) to enter next step

Obs	newvar	alpha	rp	cr	ratio
1	source	0.596	0.011	0.019	0.596
2	pow	0.000	0.000	0.006	0.000
3	ms	0.354	0.005	0.014	0.354
4	psx	0.532	0.009	0.016	0.532
5	aor	0.231	0.002	0.007	0.231
6	homstate	0.467	0.041	0.087	0.467
7	means	0.337	0.005	0.015	0.337
8	nsurg	0.314	0.003	0.009	0.314
9	nbs	0.463	0.005	0.012	0.463
10	cp	0.344	0.006	0.017	0.344
11	lsr	0.417	0.008	0.020	0.417
12	ethnic	0.524	0.010	0.018	0.524
13	race1	0.358	0.007	0.018	0.358
14	psrcd	0.455	0.006	0.013	0.455

The Forward method continues to Step 8, selecting the variable *nbs*, and stopped at Step 9 as including the variable *source* would have increased the value of RP to 34.6% which is larger than the 30% stopping criterion. Notice that the value of α for the variable *source* in Step 9 is computed by (1), 1.078 = 0.934 / 0.866, where 0.866 is the values of $\frac{RP}{CR}$ as the variable *nbs* has been entered in Step 8 and 0.934 is the value of $\frac{RP}{CR}$ if the variable *source* is to be entered in Step 9.

Forward step 8:

Varlist: sex ag8r pow aor nsurg psrcd psx means cp
nbs (smallest alpha=1.091) to enter next step

Obs	newvar	alpha	rp	cr	ratio
1	source	1.136	0.251	0.278	0.902
2	ms	1.134	0.281	0.312	0.901
3	homstate	1.300	0.685	0.664	1.032
4	nbs	1.091	0.220	0.254	0.866
5	lsr	1.135	0.330	0.366	0.901
6	ethnic	1.122	0.253	0.284	0.891
7	race1	1.146	0.288	0.316	0.910

Forward step 9:

Varlist: sex ag8r pow aor nsurg psrcd psx means cp nbs
source (smallest alpha=1.078) to enter next step
Stopped at step 9, 30% criterion met

Obs	newvar	alpha	rp	cr	ratio
1	source	1.078	0.346	0.370	0.934
2	ms	1.086	0.379	0.403	0.941
3	homstate	1.201	0.767	0.737	1.040
4	lsr	1.101	0.425	0.445	0.954
5	ethnic	1.094	0.353	0.372	0.948
6	race1	1.100	0.385	0.404	0.953

The summary of Forward method shows how the set of variables is built up and the values of RP are increased over the steps.

Summary of Forward Selection method
Stopped when pp > 30%

Obs	step	varlist	alpha	rp	cr	ratio
1	F1	sex ag8r pow	0.000	0.000	0.006	0.000
2	F2	sex ag8r pow aor	.	0.003	0.013	0.224
3	F3	sex ag8r pow aor nsurg	1.819	0.012	0.030	0.407
4	F4	sex ag8r pow aor nsurg psrcd	1.466	0.030	0.050	0.597
5	F5	sex ag8r pow aor nsurg psrcd psx	1.078	0.034	0.054	0.643
6	F6	sex ag8r pow aor nsurg psrcd psx means	1.138	0.079	0.108	0.732
7	F7	sex ag8r pow aor nsurg psrcd psx means cp	1.086	0.138	0.174	0.794
8	F8	sex ag8r pow aor nsurg psrcd psx means cp nbs	1.091	0.220	0.254	0.866

In Step 1 of Backward method, the variable *psx* with the largest α = 0.991 is removed.

Backward step 1:

Varlist: sex ag8r source pow ms psx aor homstate means nsurg nbs cp lsr ethnic race1 psrcd
psx (largest alpha=0.991) to be removed next step

Obs	delvar	alpha	rp	cr	ratio
1	source	0.987	0.996	0.983	1.013
2	pow	0.987	0.998	0.985	1.013
3	ms	0.982	0.994	0.976	1.018
4	psx	0.991	0.999	0.990	1.009
5	aor	0.990	0.998	0.988	1.010
6	homstate	0.968	0.931	0.901	1.034
7	means	0.990	0.998	0.987	1.011
8	nsurg	0.990	0.998	0.988	1.011
9	nbs	0.990	0.998	0.988	1.010
10	cp	0.988	0.998	0.986	1.012
11	lsr	0.980	0.988	0.969	1.020
12	ethnic	0.987	0.997	0.984	1.014
13	race1	0.988	0.997	0.985	1.012
14	psrcd	0.991	0.999	0.990	1.009

The Backward method stopped at Step 12 as removing the variable *ms* with the largest α = 1.57 would have caused the value of RP lower than the stopping criterion of 5%. According to (1), the value of α is computed as 1.57 = 0.633 / 0.403, where 0.633 is the value of $\frac{RP}{CR}$ as the variable *ethnic* was removed in Step 11 and the value of $\frac{RP}{CR}$ would be 0.403 if the variable *ms* is to be removed in Step 12.

Backward step 11:

Varlist: sex ag8r pow ms cp ethnic
ethnic (largest alpha=1.287) to be removed next step

Obs	delvar	alpha	rp	cr	ratio
1	pow	1.051	0.101	0.130	0.775
2	ms	1.150	0.064	0.090	0.708
3	cp	1.090	0.060	0.081	0.747
4	ethnic	1.287	0.058	0.092	0.633

Backward step 12:

Varlist: sex ag8r pow ms cp
ms (largest alpha=1.570) to be removed next step
Stopped at step 12, 5% criterion met

Obs	delvar	alpha	rp	cr	ratio
1	pow	1.168	0.030	0.056	0.542
2	ms	1.570	0.013	0.031	0.403
3	cp	1.427	0.012	0.026	0.444

The summary of Backward method shows how the variables are dropped and the values of RP are decreased to a value just above a pre-set value of 5% over the steps.

Summary of Backward Selection method
Stopped when pp < 5%

Obs	step	varlist
1	B1	sex ag8r source pow ms aor homstate means nsurg nbs cp lsr ethnic race1 psrcd
2	B2	sex ag8r source pow ms aor homstate means nsurg cp lsr ethnic race1 psrcd
3	B3	sex ag8r source pow ms aor homstate means nsurg cp lsr ethnic race1
4	B4	sex ag8r source pow ms aor homstate nsurg cp lsr ethnic race1
5	B5	sex ag8r source pow ms homstate nsurg cp lsr ethnic race1
6	B6	sex ag8r source pow ms homstate cp lsr ethnic race1
7	B7	sex ag8r source pow ms cp lsr ethnic race1
8	B8	sex ag8r source pow ms cp ethnic race1
9	B9	sex ag8r source pow ms cp ethnic
10	B10	sex ag8r pow ms cp ethnic
11	B11	sex ag8r pow ms cp

Obs	alpha	rp	cr	ratio
1	0.991	0.999	0.990	1.009
2	0.999	0.998	0.988	1.010
3	0.999	0.997	0.986	1.011
4	0.996	0.994	0.980	1.015
5	0.995	0.989	0.969	1.020
6	0.993	0.984	0.957	1.027
7	1.004	0.708	0.692	1.023
8	1.081	0.434	0.459	0.946
9	1.086	0.262	0.301	0.871
10	1.069	0.153	0.188	0.815
11	1.287	0.058	0.092	0.633

The Stepwise method is similar to the Forward method up to Step 10. Then, it detects the variable *cp* needs to be removed (RP=0.404 > 0.35) in the Backward Step 1, and the Backward process was stopped as the variable *source* cannot be removed (RP=0.275 < 0.35) in the Backward Step 2. The Stepwise method would continue with the Forward process but stopped in Step 11 because the variable *cp* selected to be entered has just been removed in the last step, even though RP=0.523 for the variables *cp* is still lower than the stopping criterion of 0.55 .

Forward step 10:

Varlist: sex ag8r pow aor nsurg psrcd psx means cp nbs source
ms (smallest alpha=1.052) to enter next step

Obs	newvar	alpha	rp	cr	ratio
1	ms	1.052	0.523	0.533	0.983
2	homstate	1.125	0.870	0.828	1.051
3	lsr	1.052	0.565	0.575	0.983
4	ethnic	1.057	0.491	0.498	0.987
5	race1	1.063	0.525	0.529	0.993

Backward step 1:

Varlist: sex ag8r pow aor nsurg psrcd psx means cp nbs source ms
cp (largest alpha=1.048) to be removed next step

Obs	delvar	alpha	rp	cr	ratio
1	pow	1.009	0.430	0.442	0.974
2	aor	1.014	0.474	0.490	0.969
3	nsurg	1.033	0.445	0.468	0.951
4	psrcd	1.000	0.523	0.533	0.983
5	psx	0.999	0.523	0.532	0.983
6	means	1.017	0.462	0.478	0.966
7	cp	1.048	0.404	0.430	0.938
8	nbs	1.040	0.418	0.442	0.945
9	source	1.044	0.379	0.403	0.941
10	ms	1.052	0.346	0.370	0.934

Backward step 2:

Varlist: sex ag8r pow aor nsurg psrcd psx means nbs source ms
source (largest alpha=1.044) to be removed next step
Stopped at step 2, 35% criterion met

Obs	delvar	alpha	rp	cr	ratio
1	pow	1.018	0.316	0.343	0.921
2	aor	1.019	0.359	0.390	0.920
3	nsurg	1.033	0.328	0.361	0.908
4	psrcd	1.000	0.404	0.430	0.938
5	psx	0.999	0.403	0.430	0.938
6	means	1.034	0.279	0.308	0.907
7	nbs	1.044	0.301	0.335	0.898
8	source	1.044	0.275	0.306	0.898
9	ms	1.056	0.247	0.277	0.888

Forward step 11:

Varlist: sex ag8r pow aor nsurg psrcd psx means nbs source ms
cp (smallest alpha=1.000) to enter next step
Stopped at step 11, cp removed last cannot be reentered

Obs	newvar	alpha	rp	cr	ratio
1	homstate	1.071	0.932	0.886	1.052
2	lsr	1.032	0.656	0.647	1.014
3	ethnic	1.007	0.564	0.570	0.990
4	race1	1.007	0.585	0.591	0.989
5	cp	1.000	0.523	0.533	0.983

Summary of Stepwise Selection method
Stopped when pp > 55%

Obs	step	varlist	alpha	rp	cr	ratio
1	F1	sex ag8r pow	0.000	0.000	0.006	0.000
2	F2	sex ag8r pow aor	.	0.003	0.013	0.224
3	F3	sex ag8r pow aor nsurg	1.819	0.012	0.030	0.407
4	F4	sex ag8r pow aor nsurg psrcd	1.466	0.030	0.050	0.597
5	F5	sex ag8r pow aor nsurg psrcd psx	1.078	0.034	0.054	0.643
6	F6	sex ag8r pow aor nsurg psrcd psx means	1.138	0.079	0.108	0.732
7	F7	sex ag8r pow aor nsurg psrcd psx means cp	1.086	0.138	0.174	0.794
8	F8	sex ag8r pow aor nsurg psrcd psx means cp nbs	1.091	0.220	0.254	0.866
9	F9	sex ag8r pow aor nsurg psrcd psx means cp nbs source	1.078	0.346	0.370	0.934
10	F10	sex ag8r pow aor nsurg psrcd psx means cp nbs source ms	1.052	0.523	0.533	0.983
11	B1	sex ag8r pow aor nsurg psrcd psx means nbs source ms	1.048	0.404	0.430	0.938

Discussion

From the examples demonstrated, the outcomes of subset of variables for release can vary by using different methods or stopping criteria. The goal of these methods is only to provide some reasonable choices of subsets of variables for release instead of finding the optimal one, which may not be necessary and is not feasible in large data. In practice, a prior knowledge or expert opinion about data should be used to determine the set of variables for disclosure first, and then let the variable selection methods to decide on those variables uncertain for release. The examples have used the cutoff of 3 for minimum cell size to be released, a change of the cutoff value will greatly affect the measurement of RP and CR and hence the selection of variables to be released. The ratio of $\frac{RP}{CR}$ is simply an increasing function of RP and a decreasing function of CR, hence other functions with a similar property maybe used as a selection criterion as well. The difference of RP-CR was not found worked well as selecting of key variables were dominated by big increases of CR instead of small increases in RP. However, the revised difference of RP – 0.5*CR worked properly in test data. Further research will be necessary determining a generally adequate selection criterion. The success of using the risk metrics of RP and CR will depend on the assumption that the target is believed coming from the sample or the sample covers the majority of population. Other choices of rick metrics are possible.

Variables pre-screening will reveal dependency among variables but not those near dependency. The variables selection methods should avoid selecting similar variables in data, as admitting similar variables may lead to difficulty computing risk measure and lead to incorrectly selecting variables in the subsequent steps.

Key Variables by Subgroup

The risk and benefit of a subset of key variables can be substantially different by subgroup in population such as male/female and rural/urban to warrant a separate analysis. More homogeneous data within a subgroup will result to better outcomes for the variables selection methods. Regrouping variables or isolating subpopulations with large

estimated odds ratios of high risk cells (with small cell size) will also prepare data better suitable for the variables selection methods.

Conclusion

The paper is to address the problem in disclosing large data for public use and avoiding risk of re-identification. To solve the problem, we gauged the usefulness of released data and its risks of being re-identified, designed the criterion of selecting variables for disclosure, and proposed systematic methods comparing subsets of variables in a step-by-step manner considering both risks and benefits. The methods will stop when the step has reached within a pre-set limit of tolerance level and then the subsets of variables in the final steps would be reasonable choices of variables for release.

Variables selection is essential for proper disclosure of data with a large number of variables. The success of methodology depends on the design of selection criterion considering both risks and benefits of released data.

Acknowledgements

Hwai-Tai Lam[1] provided test data and feedback of using the programs of the methods on real data. George Fitzelle[1] described government regulations on privacy protection. Lisa Mavrogianis[1] provided information on the Open Data policy.

References

[1] Susan Hickey, Managing Re-identification Risk for Patient-Level Datasets, June 12, 2013
[2] Susan Hickey, Protecting VHA Patient Privacy When Releasing Information, July 25, 2013
[3] XH Andrew Zhou et al, Risk of Linking HIPAA De-identified Rheumatoid Arthritis Research Dataset with CMS Data, Dec 21, 2012
[4] Khaled El Emam, Methods for the De-identification of Electronic Health Records for Genomic Research, Appendix: Measuring the Probability of Re-identification, Genome Medicine, 2011
[5] Sweeney, L., Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3, Pittsburgh 2000.
[6] Informative Graphics Corp., Redaction in Health Information Management, an IGC White Paper, Scottsdale AZ, 2013